Journal of Medical Microbiology & Diagnosis

Weltman, J Med Microb Diagn 2014, 3:2
DOI: 10.4172/2161-0703.1000135

Research Article Open Access

# An Immunobioinformatic Comparison of Influenza A Subtype Hemagglutinins

**Joel K Weltman***

*Department of Medicine, Alpert Medical School, Brown University, USA*

**Abstract**

The purpose of this research is to identify nucleotide and amino acid positions are essential for the function of the HA gene and its protein products. A metric for sequence variability was Hamming distance, determined for all possible inter-subtype HA sequence pairs of H1N1, H3N2 and H6N1 HA sequences in the complete NCBI HA gene datasets. Almost all (97.22%) of nucleotide positions at which the Hamming distance was zero were in the HA2 domain of the HA gene; the invariant nucleotides occupied second and first codon positions except for a tail-region encoded invariant tryptophan. In contrast with the results at the nucleotide level, the patterns of epitope distribution in the encoded HA proteins were similar except for a 25-amino acid sequence (283-307) in the HA1 region of HA H6N1. These results demonstrate the occurrence of similar organization of immunological epitopic biopatterns in influenza A hemagglutinins at the protein level, even in the face of large differences in the sequences of encoding nucleotides and encoded amino acids.

**Keywords:** Influenza A; Subtypes; Hemagglutinin; Bioinformatics; Hamming distance; Bepipred; Epitopes

## Introduction

The hemagglutinin (HA) gene of the influenza A virus encodes proteins that enable the influenza virus to bind to sialic acid on target cell membranes and to be internalized by those cells through a process involving fusion of the cell lipid membrane with the viral lipid membrane [1,2]. After syntheses of the hemagglutinin protein (HA0) encoded by the HA gene, the HA0 protein is cleaved by cellular proteases into a signal peptide, an HA1 protein and an HA2 protein [3]. The HA1 protein contains the recognition site for binding the virus to the sialic acid [4,5] on the membrane of the target cell. Internalization of the virus into the target cell takes place by a process of membrane fusion, mediated by the HA2 protein [1]. The influenza virus thus enters into the endosomes of the cells where, following lowering of pH, the virus replicates. The N-terminal signal peptide of the HA0 protein is involved with translocation of the nascent influenza virus across the endoplasmic reticulum membrane of the infected cell [6]. Other regions encoded by the HA gene, mainly in the HA2 domain, are involved with packaging of the (-) RNA and the proteins of the virus into the functional viral structure [7] although participation of HA2 may not be critical in the packaging process [8]. Immunogenic regions of the HA1 protein, involved in binding the virus to the host cell surface are important targets in vaccine development [9,10].

A purpose of this research is to help identify patterns of organization in the nucleotide positions of the influenza A HA gene and the HA proteins. A reduced variability at such positions may reflect biological constraints on that variability, especially given the high mutation rate of influenza viruses [11] and the lack of error-correction [12]. Nucleotide variability at second codon positions reflects variability of protein sequence since there are no degenerate codons with mutations in the second position [13].

In this research on sequence variation in the influenza A virus, the HA genes of subtypes H1N1, H3N2 and H6N1 are analyzed. Subtypes H1N1 and H3N2 are the causes of epidemics and pandemics [3], while H6N1 has posed a recent threat [14]. The research presented here uses Hamming distance [15,16] as the metric for comparing the HA gene nucleotide variation. Hamming distance does not require prior realignment of the sequences according to statistical criteria and is defined only for sequences of equal length. Hamming distance has been used to analyze the HA gene of individual HA subtypes [17,18] but, to my knowledge, this is the first report of the simultaneous application of Hamming distance to HA genes of multiple subtypes. The second codon Hamming distance distributions reflect differences in protein sequence. Despite the differences in protein sequences, the patterns of B-cell linear epitope distributions remained intact in the encoded HA proteins. Thus, it is reported here that nucleotide variation in regions of the HA gene is non-randomly distributed and it is proposed that these observed variational patterns reflect host immunological and other biologically significant evolutionary forces that act on the HA proteins.

## Methods

Sequence sorting, codon translation, data analysis and plotting were performed with Python 2.7 (EPD 7.3-1, 64-bit), Numpy 1.6.1, Scipy (0.10.1) and Biopython 1.61. Consensus sequences were determined with JalView [19]. Hamming distances [15,16] were computed with Python 2.7.3 on the computer facilities of the Brown University Center for Computation and Visualization (CCV) for all possible inter-subtype sequence pairs. Hamming distances at first, second and third codon positions were obtained by decimative multiplication, as previously reported for arrays of information entropy [20].

The entire set of FLU project influenza A HA coding region nucleotide sequences (17017) was downloaded from the NCBI Influenza Virus Resource Database [21] in FASTA format [22] on December 26, 2013 along with an additional special set of H7N9 HA sequences linked through the NCBI website. These nucleotide

***Corresponding author:** Joel K Weltman, Department of Medicine, Alpert Medical School, Brown University, USA, Tel: 401/245-7588; E-mail: joel_weltman@brown.edu

**Received** June 25, 2013; **Accepted** June 29, 2014; **Published** July 01, 2014

**Citation:** Weltman JK (2014) An Immunobioinformatic Comparison of Influenza A Subtype Hemagglutinins. J Med Microb Diagn 3: 135. doi:10.4172/2161-0703.1000135

sequences are DNA versions of the influenza virus HA mRNA. Only complete sequences were downloaded and laboratory strains were excluded. The hemagglutinin (HA) gene sequences of the following influenza A virus subtypes were sorted from the complete download set: H1N1, H2N2, H3N2, H5N1, H6N1 and H7N9. It was determined that 82.85% of the H1N1 HA sequence dataset and 99.77% of the H3N2 HA gene dataset consisted of complete HA gene nucleotide segments of length 1701, with 16.63% of the H1N1 HA dataset being of length 1698 nucleotides. One of the H6N1 sequences was of length 1698 and one was of length 1704; the remaining H6N1 sequences (98.77%) were of length 1701. The HA genes of none of the other influenza A subtypes were 1701 nucleotides in length. Thus, the HA genes of influenza A subtypes H1N1, H3N2 and H6N1 of length 1701 nucleotides provided the requisite and suitable sequence datasets for a Hamming distance-based analysis.

The entire set of FLU project influenza A HA protein sequences (13944) was downloaded from the NCBI Influenza Virus Resource Database [21] in FASTA format [22] on May 27, 2014, yielding 5905 H1N1 intact HA protein sequences (yield=96.79%), 5268 H3N2 intact HA protein sequences (yield = 98.39%) and 184 H6N1 intact HA protein sequences (yield = 99.46%).

Hamming distance (***d***) at a single nucleotide position of two nucleotide sequences (***i, j)*** of equal length, one sequence from set ***i*** and one sequence from set ***j,*** is defined as ***d=0*** if ***nucleotide(i)=nucleotide(j)*** and as ***d=1*** if ***nucleotide(i) ≠ nucleotide(j)***:

$$\boldsymbol{d} = \begin{cases} 0, & nucleotide(i) = nucleotide(j) \\ 1, & nucleotide(i) \neq nucleotide(j) \end{cases} \quad (1)$$

The total Hamming distance (***D***) at a single nucleotide position of the HA subtype sets is defined as:

$$D = \sum_{\Phi} d \quad (2)$$

where ***d is*** summed over φ**,** and φ is the total number of possible pairs between sequences ***i*** and ***j.*** The summed Hamming distance (**ΣD**) between paired ***(i, j)*** HA subtype gene sequences is:

$$\sum \boldsymbol{D} = \sum_{n} \boldsymbol{D}(\boldsymbol{HA}_1, \boldsymbol{HA}_2) \quad (3)$$

where **D** is summed over ***n*** nucleotide positions. In the HA genes of the three subtypes analyzed in this study, the maximum value of ***n*** is ***L***, where ***L*** is the length of the complete H1N1, H3N2 and H6N1 HA genes and ***L***=1701.

Prediction and detection of B-cell linear epitopes were performed on the influenza A consensus HA protein sequences with Bepipred [23] using the recommended default cutoff score of 0.35. The Bepipred website was accessed via the Immune Epitope Database and Analysis Resource, La Jolla Institute for Allergy & Immunology, CA (http://www.iedb.org/). Percentage identity (PID) was calculated for the raw concensus protein sequences and for consensus sequences that were realigned with Clustal-Omega [24; http://www.ebi.ac.uk/Tools/services/web_clustalo/toolform.ebi].

## Results

FLU Project HA sequences of length 1701 nucleotides of the following subtypes (sequence numbers in parentheses) were downloaded from the NCBI Influenza Virus Resource database: H1N1 (5655), H3N2 (4107) and H6N1 (160). In order to remove sequences that could interfere with Hamming distance analysis, downloaded sequences which could not serve as templates for translation into full-length (566 amino acid) HA0 proteins were purged, yielding: H1N1 (5365; 94.87% of download), H3N2 (3989; 97.12% of download) and H6N1 (159; 99.38% of download). The (H1N1, H3N2) Hamming distance, summed over 1701 positions, was computed from 5365*3989=21,400,985 sequence pairs, the (H1N1, H6N1) Hamming distance was computed from 5365*159=853, 035 sequence pairs and the (H3N2, H6N1) Hamming distance was computed from 3989*159=634, 251 sequence pairs. Hamming distances between the pairs of these subtype sequences are shown in Table 1. The median Hamming distance of 1223 observed between HA genes of subtypes H1N1 and H3N2 represents a frequency 0.7190 summed over the 1701 nucleotide positions. Using that frequency as a probability for a binomial approximation leads to a predicted standard deviation that is 2.54 greater than the observed standard deviation. Similarly, the standard deviations binomially predicted for the (H1N1, H6N1) pairs and for the (H3N2, H6N1) pairs are 3.20 and 2.75 greater than those observed, respectively. The summed Hamming distance of 883 for the (H1N1, H6N1) paired datasets is less than those observed for both the (H1N1, H3N2) paired datasets and for the (H3N2, H6N1) paired datasets; there was no overlap between the (H1N1, H6N1) Hamming distance range with those of the other two dataset pairs. The probability (**p**) values for the observed absolute non-overlap between the (H1N1, H6N1) Hamming distance and the Hamming distances of the (H1N1, H3N2) and (H3N2, H6N1) datasets are: $\mathbf{p}<(1/853{,}035)*(1/21{,}400{,}985.)=5.4777\times10^{-14}$ and $\mathbf{p}<(1/853{,}035.)*(1./634{,}251.)=1.8483 \times 10^{-12}$, respectively.

The cumulative distributions of Hamming distances (**Σ*D***) over the complete lengths of the H1N1, H3N2 and H6N1 HA genes are shown in Figure 1. Hamming distance is distributed linearly between HA(H1N1, H3N2) sequence pairs (Figure 1a) and between HA(H3N2, H6N1) sequence pairs (Figure 1c) but has a clearly nonlinear distribution in HA(H1N1, H6N1) sequence pairs (Figure 1b). HA(H1N1, H3N2) gene sequence pairs and HA(H3N2, H6N1) pairs are well described as linear summations of Bernoulli binomial functions [25] but correlation of a binomial function with the observed summed Hamming distance along gene length in HA (H1N1, H6N1) pairs has a large residual. The nonlinearity in Hamming distance distribution along the length of the HA genes occurs mainly in first (decimative 100) and second (decimative 010) codon positions (Figures 2), beginning at about nucleotide number 800.

Percent identity (PID) matrices are given in Table 2 for the native and for the realigned H1N1, H3N2 and H6N1 HA0 proteins. Considerable differences among the HA0 sequences are present even after realignment, especially between HA0 H3N2 and the HA0 proteins of the other two subtypes.

| | Median | Mean | Std | Minimum | Maximum |
|---|---|---|---|---|---|
| **∑D(H1N1, H3N2)** | | | | | |
| **Observed** | 1223 | 1223.06 | 7.2033 | 1183 | 1268 |
| **Predicted** | 1223 | 1223.01 | 18.536 | 1130 | 1305 |
| **∑D(H1N1, H6N1)** | | | | | |
| **Observed** | 883 | 884.128 | 6.4552 | 861 | 932 |
| **Predicted** | 884 | 884.109 | 20.6254 | 785 | 984 |
| **∑D(H3N2, H6N1)** | | | | | |
| **Observed** | 1243 | 1243.92 | 6.6514 | 1216 | 1277 |
| **Predicted** | 1244 | 1243.92 | 18.2835 | 1157 | 1331 |

**Table 1:** Hamming Distances Summed Over the 1701 Nucleotide Positions of the Hemagglutinin Gene. Influenza A subtypes of the paired gene sequences of the sequence datasets are identified in parentheses. The predicted values were obtained from 1 x $10^6$ pseudorandom Bernoulli binomial trials.

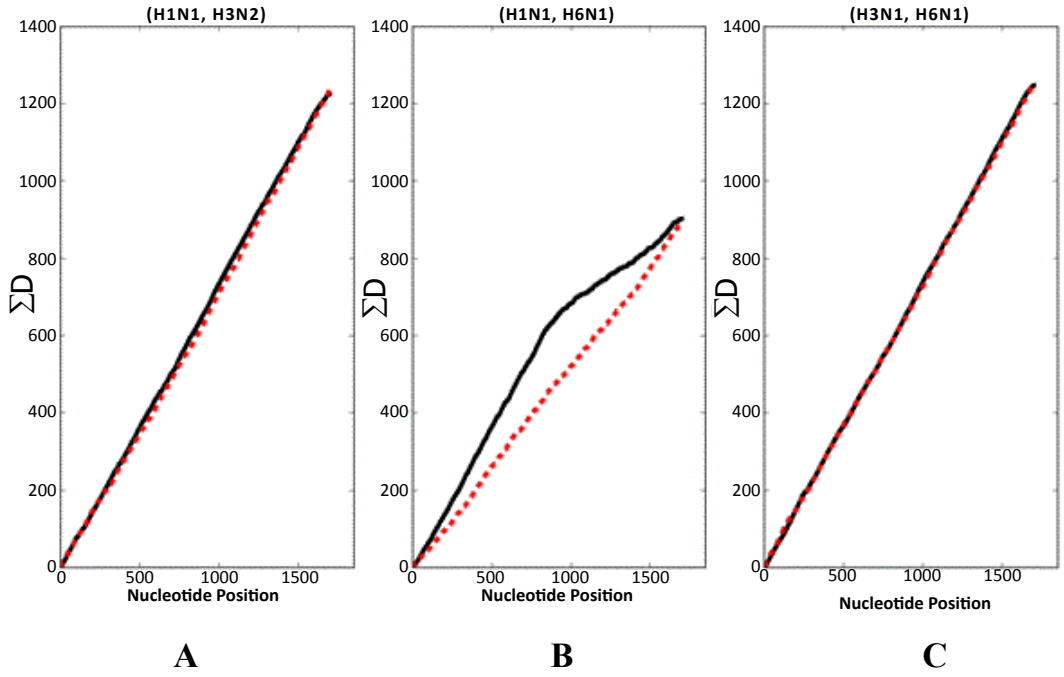


**Figure 1:** Cumulative Hamming Distances ($\Sigma D$) Between Paired Influenza A Subtype Hemagglutinins. A. Data are shown for (H1N1, H3N2) sequence pairs (left), B. (H1N1, H6N1) sequence pairs (middle) and C. (H3N2, H6N1) sequence pairs (right). Black lines (observed $\Sigma D$); red dashed lines (approximations by summed single binomial function).

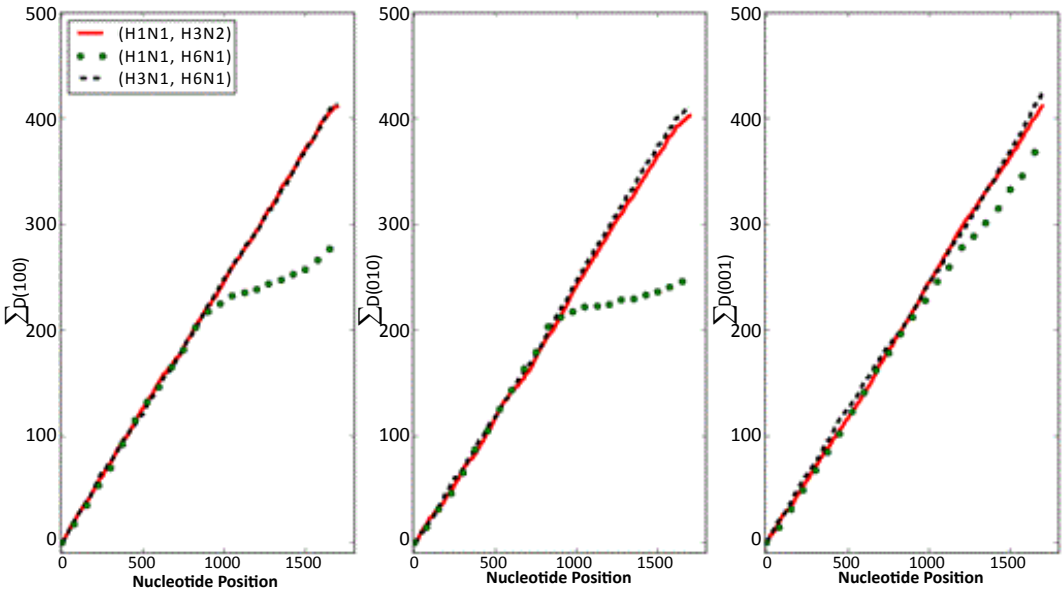


**Figure 2:** Decimative Cumulative Hamming Distances ($\Sigma D$) Between Paired Influenza A Subtype Hemagglutinins. Data are shown for (H1N1, H3N2) sequences pairs, (H1N1, H6N1) sequence pairs and (H3N3, H6N1) sequence pairs for decimative frame 100 (first codon position; left), decimative frame 010 (second codon position; center) and decimative frame 001 (third codon position; right).

Correlations between the Bepipred Scores (**B**) of the 566 amino acid positions of all pairs of HA proteins are shown in Figure 3. The Spearman correlation coefficients (r) for **B** values of the amino acid positions of the HA proteins shown in Figure 3 are : HA(H1N1) and HA(H3N2) r=0.4497, p=1.5971e-29; HA(H1N1) and HA(H6N1) r=0.8282, p=6.2993e-144; HA(H3N2) and HA(H6N1) r=0.4287, p=1.0401e-26. All of these correlations are statistically significant.

The distributions of Bepipred scores (**B**) for each of the three

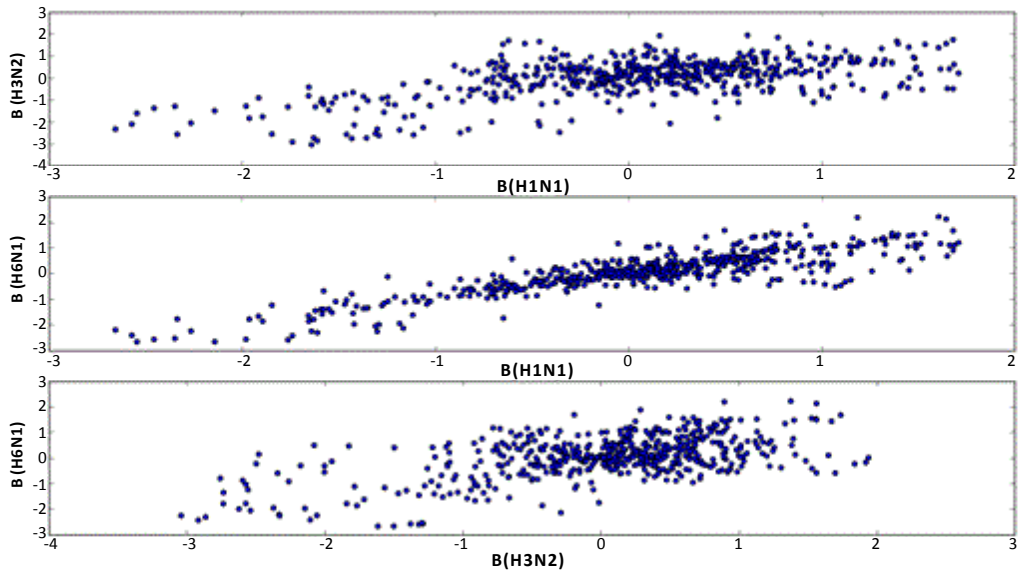


**Figure 3:** Correlations between Influenza A Subtype HA Consensus Sequence Bepipred Scores (B) for Each Amino Acid Position of HA Subtypes H1N1, H3N2 and H6N1. (top) x=H1N1, y=H3N2; (middle) x=H1N1, y=H6N1 (bottom) x=H3N2, y=H6N1.

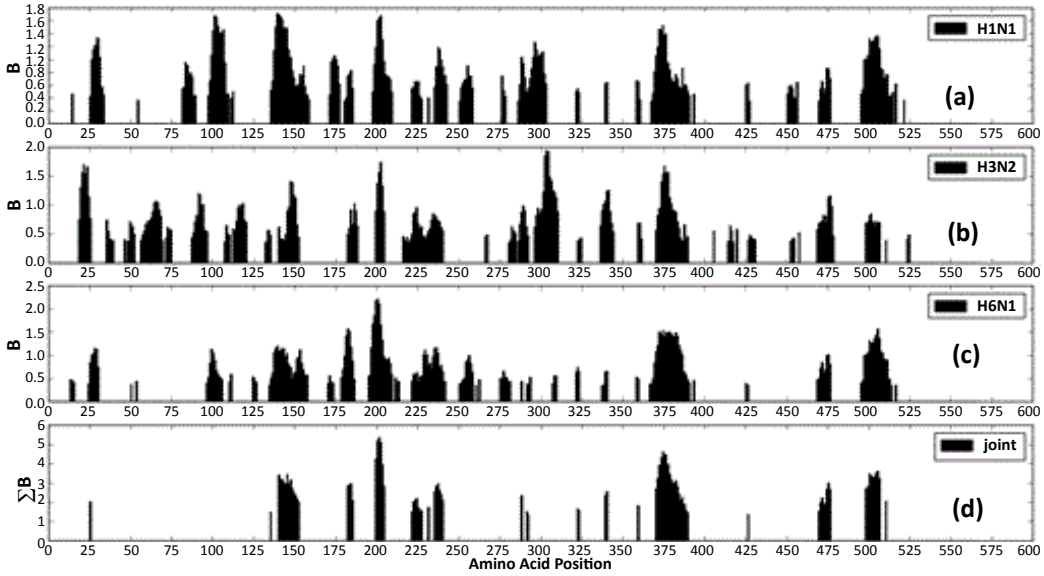


**Figure 4:** Bepipred Scores (**B**) for Amino Acid Positions of Influenza A Subtype HA Proteins. **B** values are shown only for those amino acid positions with a **B** value above the cutoff (**B**> 0.35). (a) HA (H1N1); (b) HA (H3N2); (c) HA(H6N1); (d) summed **B** scores (ΣB) for positions that are significant (>0.35) jointly for all three subtypes.

influenza A subtype HA proteins are shown in Figure 4. Only values above the recommended cutoff (0.35) are shown. **B** values at positions that were significant for all three HA subtypes were summed and are also shown for reference and comparison. There were 86 amino acid positions with significant **B** scores jointly, in all three subtypes. Clusters of amino acids at positions with significant **B** values were observed at HA protein positions 141-153, 200-205, 222-228, 236-241 and 371-390. Other, smaller joint epitope clusters were distributed throughout the sequences. There was a paucity of epitopic residues in amino acid residues 283-307 of the H6N1 HA protein.

## Discussion

A goal of this research is to detect nonrandom patterns in the variation of the HA gene and protein of subtypes H1N1, H3N2 and

| | H1N1 | H3N2 | H6N1 |
|---|---|---|---|
| **H1N1** | 100.00 | 7.42 | 41.52 |
| **H3N2** | 7.42 | 100.00 | 6.36 |
| **H6N1** | 41.52 | 6.36 | 100.00 |

(a) Native HA0 sequences

| | H1N1 | H3N2 | H6N1 |
|---|---|---|---|
| **H1N1** | 100.00 | 41.52 | 59.72 |
| **H3N2** | 41.52 | 100.00 | 39.93 |
| **H6N1** | 59.72 | 39.93 | 100.00 |

(b) Realigned HA0 sequences

**Table 2:** Percent Identity (PID) Matrices for HA0 Proteins of Influences A Subtypes H1N1, H3N2 and H6N1. (a) Native, unrealigned sequences (b) realigned sequences.

| | N(nt) | N(aa) | Codon Position | Nucleotide | HA0 Domain |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 100 | A | signal peptide |
| **2** | 2 | 1 | 10 | U | signal peptide |
| **3** | 3 | 1 | 1 | G | signal peptide |
| **4** | 848 | 283 | 10 | U | HA1 |
| **5** | 1040 | 347 | 10 | U | HA2 |
| **6** | 1054 | 352 | 100 | G | HA2 |
| **7** | 1061 | 354 | 10 | U | HA2 |
| **8** | 1066 | 356 | 100 | G | HA2 |
| **9** | 1073 | 358 | 10 | G | HA2 |
| **10** | 1078 | 360 | 100 | G | HA2 |
| **11** | 1085 | 362 | 10 | U | HA2 |
| **12** | 1094 | 365 | 10 | G | HA2 |
| **13** | 1096 | 366 | 100 | U | HA2 |
| **14** | 1123 | 375 | 100 | G | HA2 |
| **15** | 1138 | 380 | 100 | G | HA2 |
| **16** | 1141 | 381 | 100 | G | HA2 |
| **17** | 1153 | 385 | 100 | A | HA2 |
| **18** | 1199 | 400 | 10 | U | HA2 |
| **19** | 1316 | 439 | 10 | A | HA2 |
| **20** | 1334 | 445 | 10 | U | HA2 |
| **21** | 1343 | 448 | 10 | A | HA2 |
| **22** | 1399 | 467 | 100 | A | HA2 |
| **23** | 1418 | 473 | 10 | A | HA2 |
| **24** | 1426 | 476 | 100 | G | HA2 |
| **25** | 1442 | 481 | 10 | G | HA2 |
| **26** | 1457 | 486 | 10 | A | HA2 |
| **27** | 1535 | 512 | 10 | U | HA2 |
| **28** | 1592 | 531 | 10 | U | HA2 |
| **29** | 1628 | 543 | 10 | U | HA2 |
| **30** | 1645 | 549 | 100 | G | HA2 |
| **31** | 1655 | 552 | 10 | U | HA2 |
| **32** | 1657 | 553 | 100 | U | HA2 |
| **33** | 1658 | 553 | 10 | G | HA2 |
| **34** | 1659 | 553 | 1 | G | HA2 |
| **35** | 1663 | 555 | 100 | U | HA2 |
| **36** | 1664 | 555 | 10 | G | HA2 |
| **37** | 1679 | 560 | 10 | U | HA2 |
| **38** | 1685 | 562 | 10 | G | HA2 |
| **39** | 1694 | 565 | 10 | G | HA2 |
| **40** | 1699 | | 100 | U | TER |

**Table 3:** HA Nucleotide Positions with Zero Hamming Distance Between All H1N1, H3N2 and H6N1 Sequence Pairs. Nucleotide (nt) and amino acid (aa) positions are identified by number (N) in the second and third columns. Codon positions are designated as decimation frame first codon position (100), second codon position (010) and third codon position (001) in the fourth column. The termination codon is designated as TER in row 40. The HA nucleotides are identified in the mRNA representation.

H6N1 influenza A viruses because such nonrandom patterns may be direct evidence of significant biological forces acting on those viruses. HA genes of subtypes H1N1, H3N2 and H6N1 are of epidemiological importance and provide the opportunity for detecting such nonrandom patterns by Hamming distance metric because these genes are of equal length (see Results). Hamming distance provided a robust, yet sensitive metric, involving determination of sequence differences between all possible inter-subtype pairs in the sequence sets. Despite the commonality of sequence length, there were nucleotide differences between many of the nucleotide position pairs. Hamming distances were nonzero at 71.90% of (H1N1, H3N2) HA sequence pairs, 46.73% of (H1N1, H6N1) HA sequence pairs and at 65.75% of the (H3N2, H6N1) HA sequence pairs (Table 1). However, Hamming distances between the HA(H1N1, H6N1) sequence pairs were lower those of other two sets of pairs so that the HA(H1N1, H6N1) Hamming distances were completely non-overlapping, disjoint with the HA(H1N1, H3N2) and HA(H3N2, H6N1) Hamming distances. The probability (**p**) of each these results having occurred randomly is only $\mathbf{p}<5.4777 \times 10^{-14}$ and $\mathbf{p}<1.8483 \times 10^{-12}$, respectively.

The lower Hamming distance between HA (H1N1, H6N1) pairs, discussed above, is associated with nonlinear distribution of Hamming distance beginning at approximately nucleotide position 800 (Figure 1b), especially at nucleotides in first and second codon positions (Figure 2). The observed reduced distribution of Hamming distances at these codon positions suggests biological constraints acting at the protein level rather than synonymous mutations at the nucleotide level. Nucleotide position 800 of the HA gene is near the 3'-end of the HA1 region of the gene.

Thirty six (36) nucleotide positions were identified at which the Hamming distance was equal to zero between all sequence pairs of the HA genes of all three influenza A subtypes (Table 3). One of the positions (848) is in the HA1 domain of the HA gene and the other 35 positions are in the HA2 domain. 61.1% of these nucleotide positions are second positions of codons and 31.1% are first positions of codons. These results are consistent with the pattern of reduced distribution of Hamming distance in (H1N1, H6N1) HA sequence pairs (Figure 2). This pattern of constrained Hamming distance suggests processes whose actions begin on the 3'-end of the HA1 domain (near nucleotide position 848) and act at the 35 positions in the HA2 domain. Such a distribution suggests that there are constraining effects on the HA0 protein, ie, before proteolytic cleavage (1) of HA0 into HA1 and HA2 proteins.

The HA2 protein is known to have several functions in the influenza virus. HA2 protein interacts with the HA1 protein so as to enable HA1 to bind to sialic acid on the membrane of the target cell [1]. HA2 participates in a membrane fusion process during internalization of influenza virus into the target cell. HA2 participates in the packaging of nascent influenza A virus particles, thereby facilitating secretion of mature virions from endosomes [7] although it should be noted that HA2 participation in packaging may not be essential [8]. The HA2 protein provides a hydrophobic virus tail that anchors the virion to the virus lipid membrane coat [26]. The nucleotide positions and corresponding amino acid positions listed in Table 2 have not yet been specifically implicated in any of these processes. It seems especially interesting that the nucleotides at positions 1657, 1658 and 1659 are uniformly U, G, and G, encoding Trp553. Trp553 is a hydrophobic amino acid located 13 amino-acid positions from the COOH-terminus of the anchoring tail region. It is postulated here that the 36 nucleotide positions with zero Hamming distance, identified in Table 3, which are

absolutely conserved in all of the HA nucleotide sequences studied, are participating in these, and perhaps other, biological processes that are essential to the influenza A virus. Insight into the mechanisms of nucleotide conservation at these 36 nucleotide positions can be increased by analyzing biological effects of base substitutions at these positions on the structure and function of influenza A viruses.

Soundararajanin et al [27] have reported epitope networking interaction patterns within realigned HA sequences and have effectively coupled immune epitope data with crystallographic analyses. The research described here is based upon analysis of HA sequences of equal length, thereby avoiding the sequence realignment step. Avoiding sequence realignment may make it possible to detect factors influencing and governing immunological epitope distribution patterns by direct, computerized application to large sequence sets in a manner analogous to that shown for the consensus sequences in Figure 4.

The nucleotide substitutions discussed above, especially at second and first codon positions represent non-degenerate substitutions that could significantly affect the biological activity of the HA0 protein and the derivative signal peptides, HA1 and HA2 proteins of the influenza A subtypes studied. Accordingly, the comparison of epitope distributions in the HA proteins of the three subtypes was undertaken. The percent identity (PID) of these HA proteins were small, even after realignment (Table 2). Despite these differences in amino acid sequence, the patterns of epitopic potential were correlated (Figure 3) and were similarly distributed (Figure 4). The paucity of epitopic sites in amino acid residues 283-307 of the H6N1 HA protein is in the HA1 region of the HA protein. The similar overall distribution of epitopic sites in the three hemagglutinins, despite the differences in amino acid sequence and composition, is consistent with the non-randomness displayed at the nucleotide level. These non-random nucleotide and amino acid patterns suggest the existence of governing networks of biological organization. Detection and analysis of such networks in sets of sequences should increase our understanding of influenza biology and may be useful for design of vaccines.

### Acknowledgment

This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

### References

1. Skehel JJ, Wiley DC (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 69: 531-569.
2. Xu R, Wilson IA (2011) Structural characterization of an early fusion intermediate of influenza virus hemagglutinin. J Virol 85: 5172-5182.
3. Bouvier NM, Palese P (2008) The biology of influenza viruses. Vaccine 26 Suppl 4: D49-53.
4. Couceiro JN, Paulson JC, Baum LG (1993) Influenza virus strains selectively recognize sialyloligosaccharides on human respiratory epithelium; the role of the host cell in selection of hemagglutinin receptor specificity. Virus Research 29: 155-165.
5. Gamblin SJ, Haire LF, Russel RJ, Stevens DJ, Xiao B, et al. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 303: 1838-1842.
6. Sekikawa K, Lai CJ (1983) Defects in functional expression of an influenza virus hemagglutinin lacking the signal peptide sequences. Proc Natl Acad Sci U S A 80: 3563-3567.
7. Marsh GA, Hatami R, Palese P (2007) Specific residues of the influenza A virus hemagglutinin viral RNA are important for efficient packaging into budding virions. J Virol 81: 9727-9736.
8. Gao Q, Chou YY, DoÄŸanay S, Vafabakhsh R, Ha T, et al. (2012) The influenza A virus PB2, PA, NP, and M segments play a pivotal role during genome packaging. J Virol 86: 7043-7051.
9. Skowronski DM, Janjua NZ, Sabaiduc S, De Serres G, Winter AL, et al. (2014) Influenza A/Subtype and B/Lineage Effectiveness Estimates for the 2011-2012 Trivalent Vaccine: Cross-Season and Cross-Lineage Protection With Unchanged Vaccine. J Infect Dis.
10. Noh JY, Kim WJ2 (2013) Influenza vaccines: unmet needs and recent developments. Infect Chemother 45: 375-386.
11. Drake JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci U S A 90: 4171-4175.
12. Steinhauer DA, Holland JJ (1986) Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. J Virol 57: 219-228.
13. Lehmann J, Libchaber A (2008) Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. RNA 14: 1264-1269.
14. Wei SH, Yang JR, Wu HS, Chang MC, Lin JS, et al. (2013) Human infection with avian influenza A H6N1 virus: an epidemiological analysis. Lancet Respir Med 1: 771-778.
15. Hamming RW (1950) Error detecting and error correcting codes. Bell System Technical J 29: 147-160.
16. Pinheiroa HP, Pinheiroa A, Sen PK (2005) Comparison of genomic sequences using the Hamming distance. J Statistical Planning Inference 130: 325–339.
17. Creanza N, Schwarz JS, Cohen JE (2010) Intraseasonal dynamics and dominant sequences in H3N2 influenza. PLoS One 5: e8544.
18. Tria F, Pompei S, Loreto V (2013) Dynamically correlated mutations drive human Influenza A evolution. Sci Rep 3: 2705.
19. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.
20. Thompson WA, Martwick A, Weltman JK (2009) Decimative multiplication of entropy arrays, with application to influenza. Entropy 11: 351-359.
21. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, et al. (2008) The influenza virus resource at the National Center for Biotechnology Information. J Virol 82: 596-601.
22. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227: 1435-1441.
23. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. Immunome Res 2: 2.
24. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. MolSystBiol 7: 539.
25. Hoel PG (1961) Introduction to Mathematical Statistics (2ndedn) Theoretical Frequency Distributions of One Variable Wiley; New York, London.
26. Skehel JJ, Waterfield MD (1975) Studies on the primary structure of the influenza virus hemagglutinin. Proc Natl Acad Sci U S A 72: 93-97.
27. Soundararajan V, Shu Zheng S, Neel Patel N (2011) Networks link antigenic and receptor-binding sites of influenza hemagglutinin: Mechanistic insight into fitter strain propagation; Sci Rep 1: 200.